

Student: Bingyang Dou Subject: CIS576 Assignment 4: Text Analysis/Visualization and Network Visualization
This is an individual assignment. You are welcome to discuss the assignment with your colleagues, but the work you turn in must be your own.

Find a “corpus” of text that is interesting to you, and perform an analysis and word cloud visualization. There are speech transcripts from both parties that might be interesting!

- a. R: Be sure to remove words that should be removed... submit the R code, word clouds, analysis supporting why certain words were removed. (15 points)

First, I decided my topic as the ‘The Important Speeches in Steve Jobs’ Life’. I want to find out the differences among his important speeches in his legendary life and find out the keys why his speeches are always attractive and successful.

Following this idea, I searched a lot of Steve Jobs’ speeches. Finally, I selected these five speeches because they are all remarkable, technically professional, and popular.

- Steve Jobs introduces the Original Macintosh - Apple Shareholder Event (1984-01-24)
- Steve Jobs introduces the Original iMac - Apple Special Event (1998-05-06)
- Steve Jobs introduces the Original iPod - Apple Special Event (2001-10-23)
- Steve Jobs introduces the Original iPhone at Macworld SF (2007-01-09)
- Steve Jobs introduces the Original iPad - Apple Special Event (2010-01-07)

I downloaded the subtitles of those speeches from YouTube and transformed them to txt file. (Using website <http://keepvid.com/> to download SRT file.) I also deleted the parts where is not the Steve Jobs’ speaking and trimmed a little bit appropriately. Although I could get whatever I want in this way, it is also have the shortcomings. That website is not perfect. It can’t recognize the mistakes. What’s more, there is not spaces between two lines. For these reasons, I took six hours on checking the subtitles and make them as good as possible.

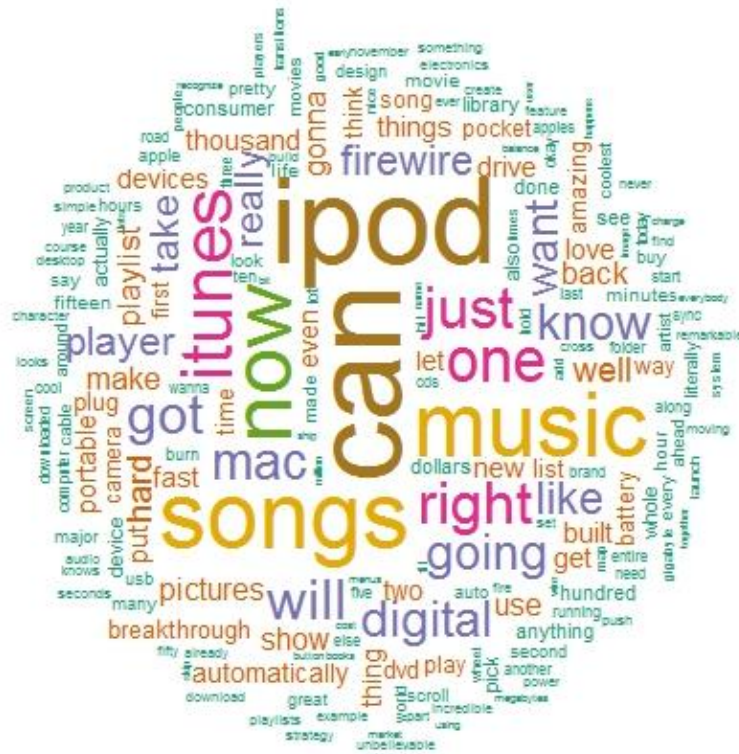
Because I generated the word files from speeches' subtitles, there are a lot of time stamps. I remove them by using 'removeNumbers' and 'removePunctuation' because the time stamps are composited by numbers and punctuation. They also remove the numbers and punctuation in the other field. Furthermore, I removed the unnecessary white space, convert everything to lower case, and remove common words by using 'stripWhitespace', 'tolower', and 'removeWords, stopwords('english')'.

I didn't remove his other words in first three speeches because I find that almost every word has its special meaning. For example, I had ever considered to remove 'thing' and 'things' in the second speech because it looks no attitude. However, finally I found that his second speech is special and the 'thing' and 'things' are special, too. I can't remove them. It is because, at that time, iMac is new product. There is no any definition about iMac. That's the way of Steve Jobs' introduction. He wants to introduce the function and the features first, leading to the concept of iMac. iMac is only the name of that thing, but that thing can realize a lot of 'was-impossible'. I removed 'just' in the fourth and the fifth speech because I can't think of any attitude or meaning about that.

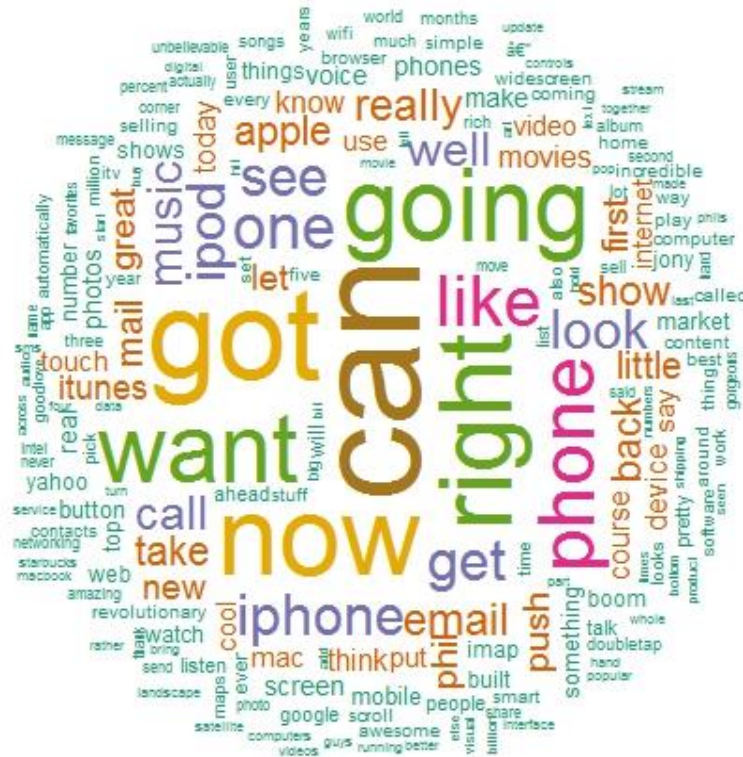
1. Macintosh



3. iPod



4. iPhone



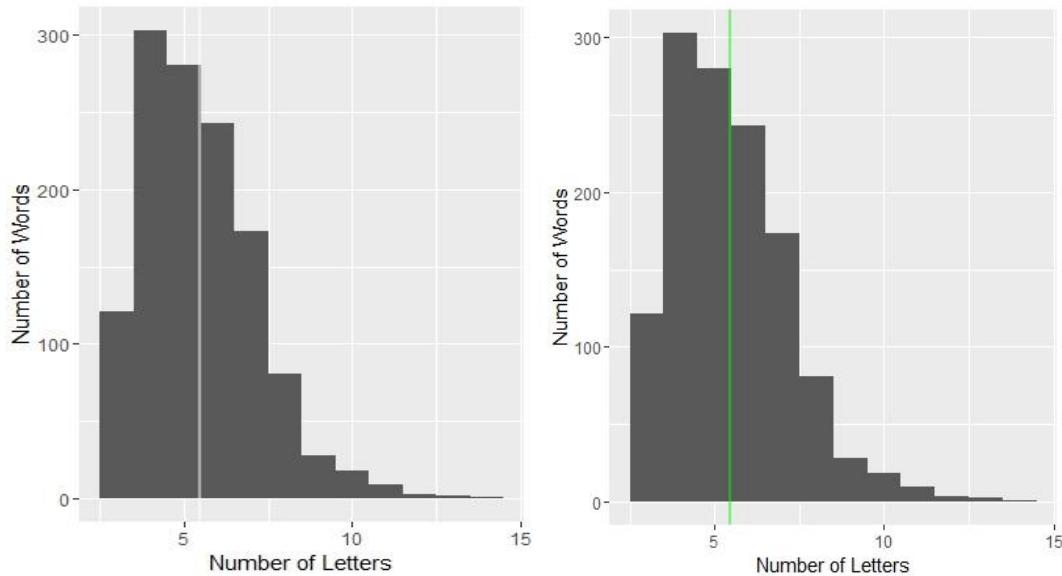
Student: Bingyang Dou Subject: CIS576 Assignment 4: Text Analysis/Visualization and Network Visualization
always trying to explain what his new products can do, what functions are new, what features are unique, and so on.

Third, he is using more and more diversity words in his speeches. From the Macintosh conference to the iPad conference, I find that there are more and more green-small words. In other words, he using more and more different words instead of focusing on the several key words, repeating again and again. In my opinion, it may be caused by two main reasons. First, there are more and more features, products are waiting him to introduce, so he had to separate his time on different words. Second, he may trying to use more and more different words to explain be he is developing his speeches' artistry. And also, because his speeches in his earlier life is time limited, the first speech may not good to represent all his personal speech habits.

Then, I find that he compared his group with IBM frequently in his first speech. I can find 'ibm' in a big size in the first speech but can't find it in the following speeches. It is easy to understand because nearly everyone knows his life is tied with ibm actually. Compare with IBM is good for apple's corporate image.

Finally, I find that he likes to introduce the milestones in timeline. We can easily find the key words about time, such as 'now', 'year', 'day', 'today'. It is very forceful to introduce the exact event by timeline. It also shows that Jobs did every steps for a long historical ambition, not for the quick money.

And then I generated the graph of 'Word Length Counts' in 1984 and 2010 because I want to find the conclusion of if Steve Jobs changed the word lengths very significantly by time.



I set the mean line in 1984 in color of white and the line in 2010 in color of green. From the graphs of 'Word Length Counts' I found that most words have about 5 to 6 letters whatever in 1984 or 2010.

By summarizing the important information of these two data. I found that the medians are always 5.000. The mean in 1984 is 5.445 and the mean of in 2010 is 5.467.

```
> summary(nchar(words))
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Data in 1984:	3.000	4.000	5.000	5.445	6.000	14.000

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Data in 2010:	3.000	4.000	5.000	5.467	6.000	14.000

I also find the tables of those two chart.

```
. table(nchar(words))
```

	3	4	5	6	7	8	9	10	11	12	13	14
Data in 1984:	61	142	147	113	75	45	14	9	1	2	1	1

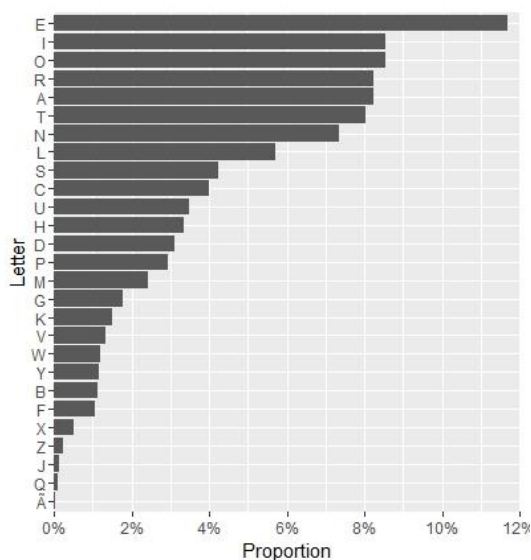
	3	4	5	6	7	8	9	10	11	12	13	14
Data in 2010:	121	303	280	243	173	81	28	18	9	3	2	1

Then I got the standard deviation in two years: in 1984 is about 2.97, in 2010 is about 1.78.

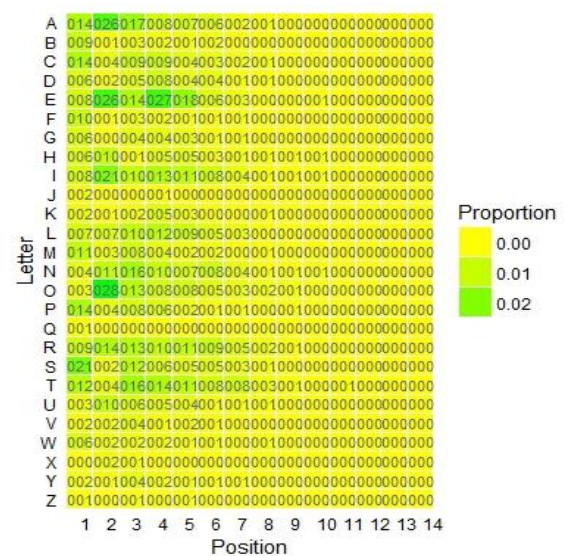
I set that $H_0: \mu = \mu_0$, $H_1: \mu \neq \mu_0$, t in t -test is about 0.005. Steve Jobs uses very similar length words in the two speeches.

Finally, I also try to build the other graphs such as 'the frequency of letters' in 2010, and 'letter and position heatmap'.

The Frequency of Letters



Letter and Position Heatmap



I found that the less frequently used letter is 'a' and the most is 'e'.

b. Tableau: Use the "cleaned" data from R. (10 points)

Prepare:

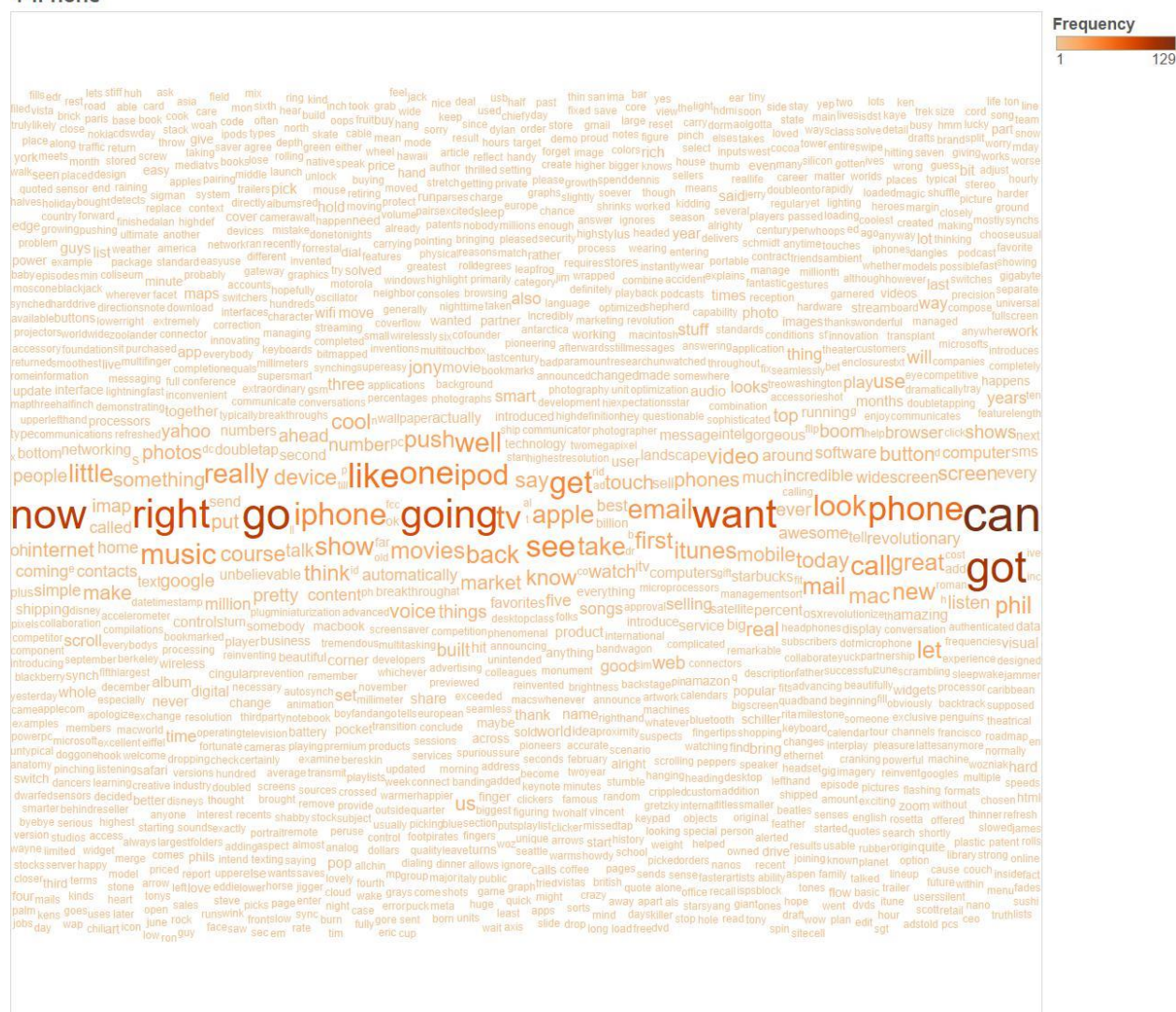
Initially, I exported all the words that have been cleaned from R. Then I downloaded the tool in this website. http://www.clearlyandsimply.com/clearly_and_simply/2015/03/the-implementation-of-word-clouds-with-excel.html I used that tool to get all the data ready.

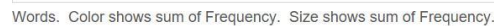
2-iMac



Words. Color shows sum of Frequency. Size shows sum of Frequency.

4-iPhone





First step I did with Tableau is generate the original speeches subtitles. By comparing

To get the overall view, I also generated two word cloud graphs include all the five

speeches. I merged all the five txt files and imported the top 30% words by the reason that

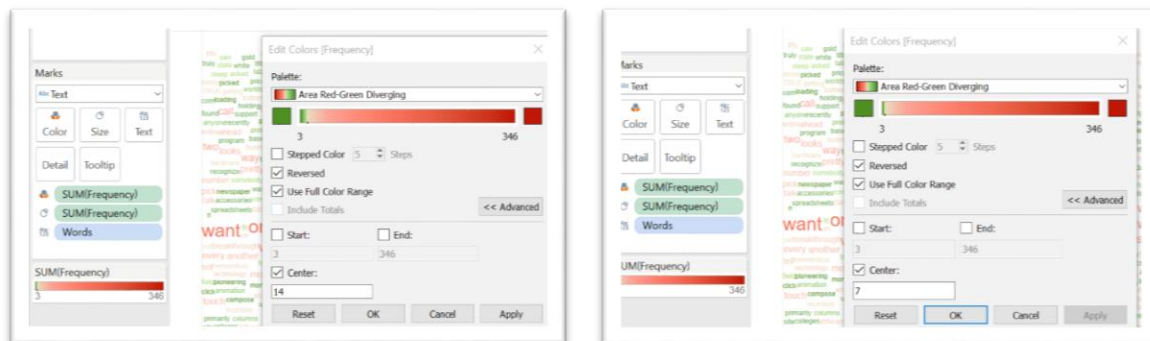
the whole file is too large to import. First, I set the color's Center at 14, which is the average

frequency of the Top 30% words. In this setting, the words' frequency lower than 14 will be

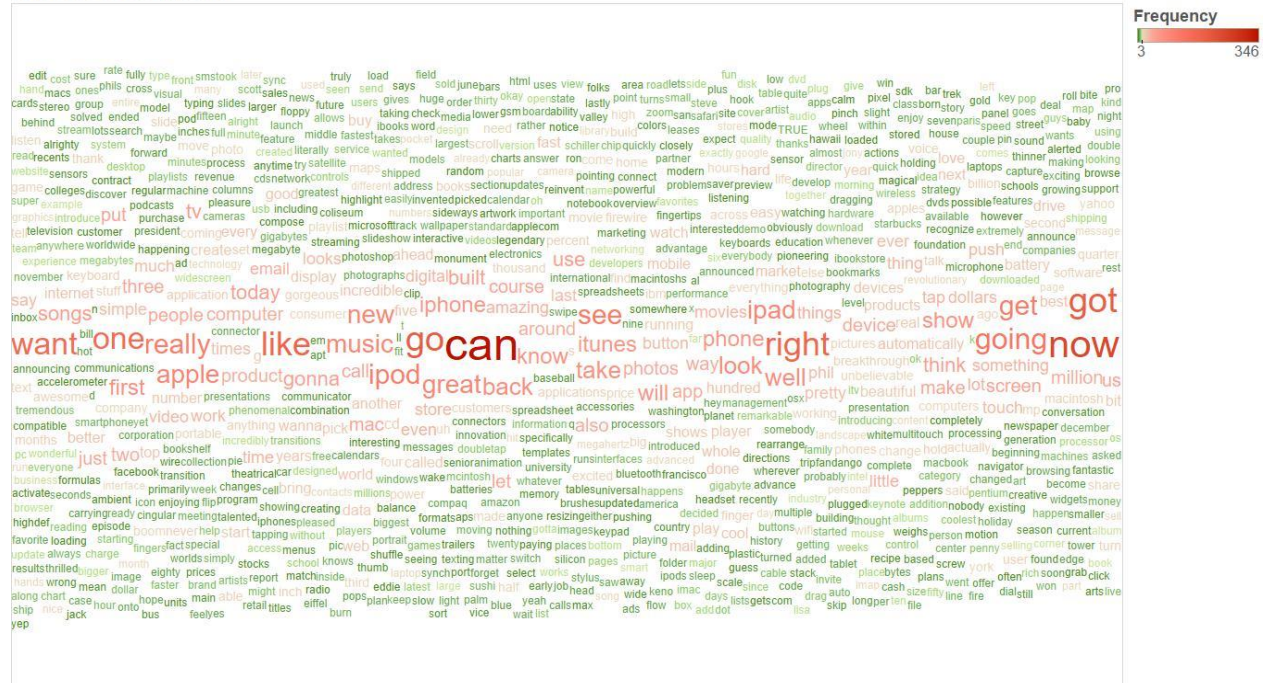
green, higher than 14 will be red. Then, I set the color's Center at 7, which is the median

frequency of the Top 30% words. In this setting, the words' frequency lower than 7 will be

green, higher than 7 will be red.

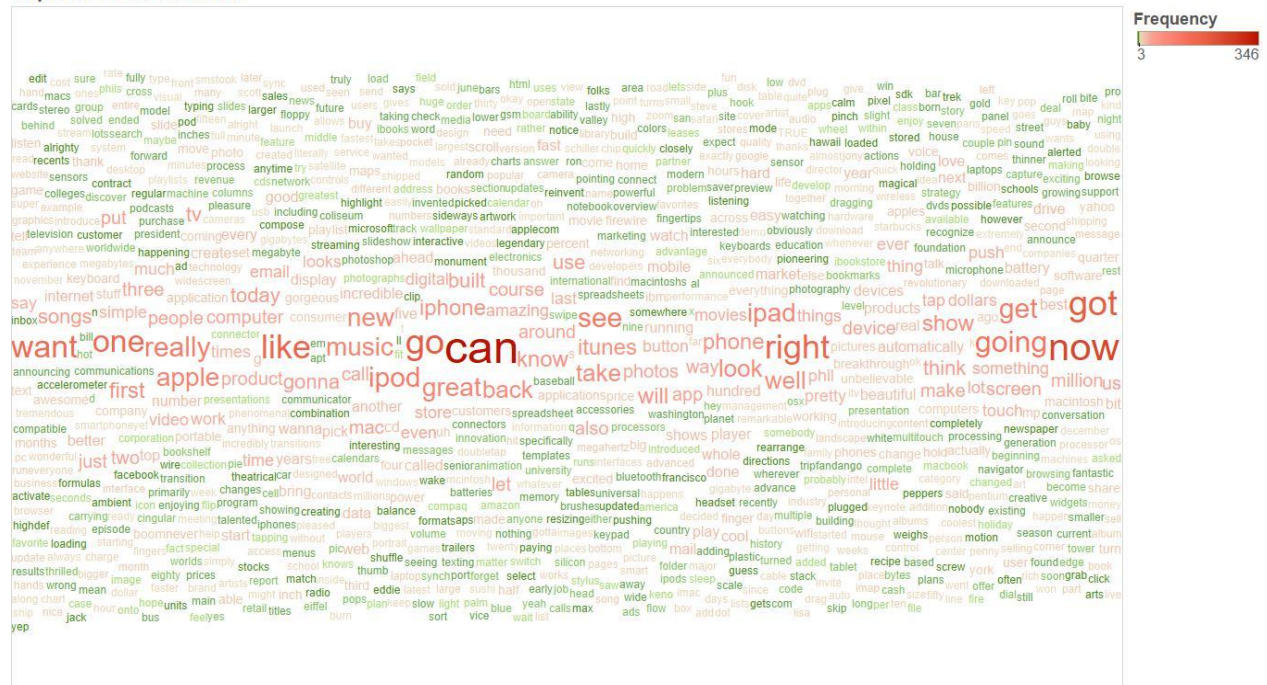


The result is shown below.



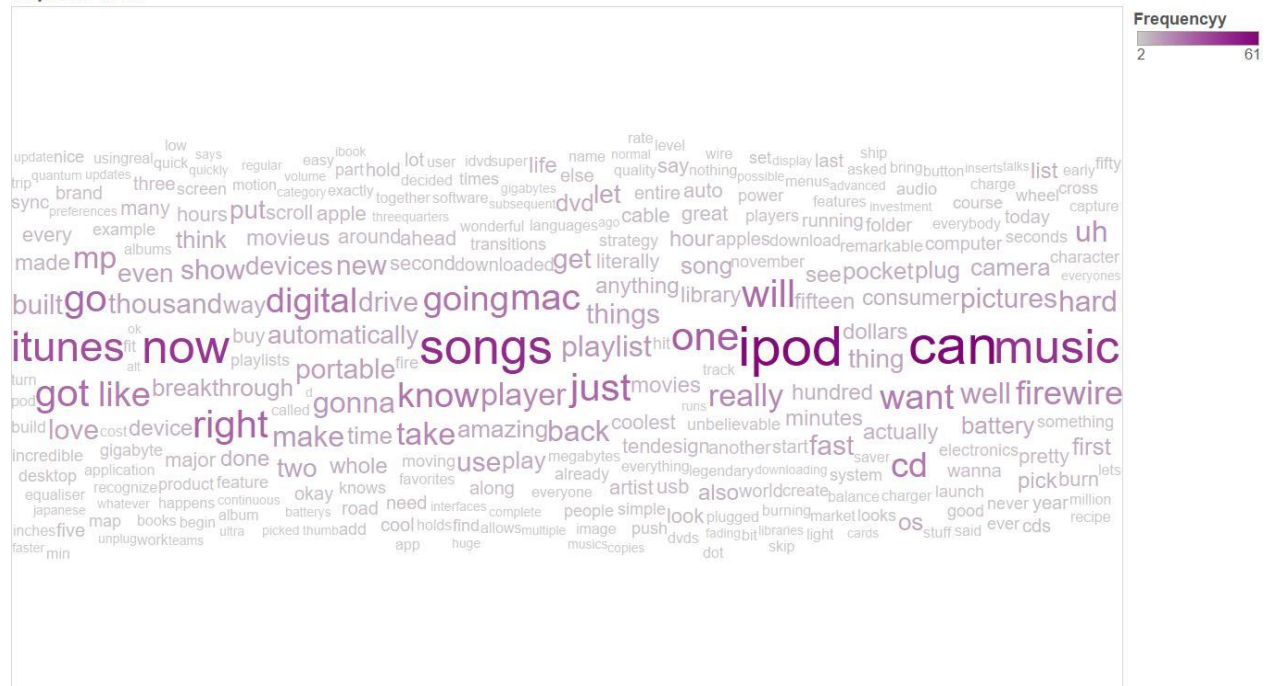
Words. Color shows sum of Frequency. Size shows sum of Frequency.

Top30%-Total-Median7



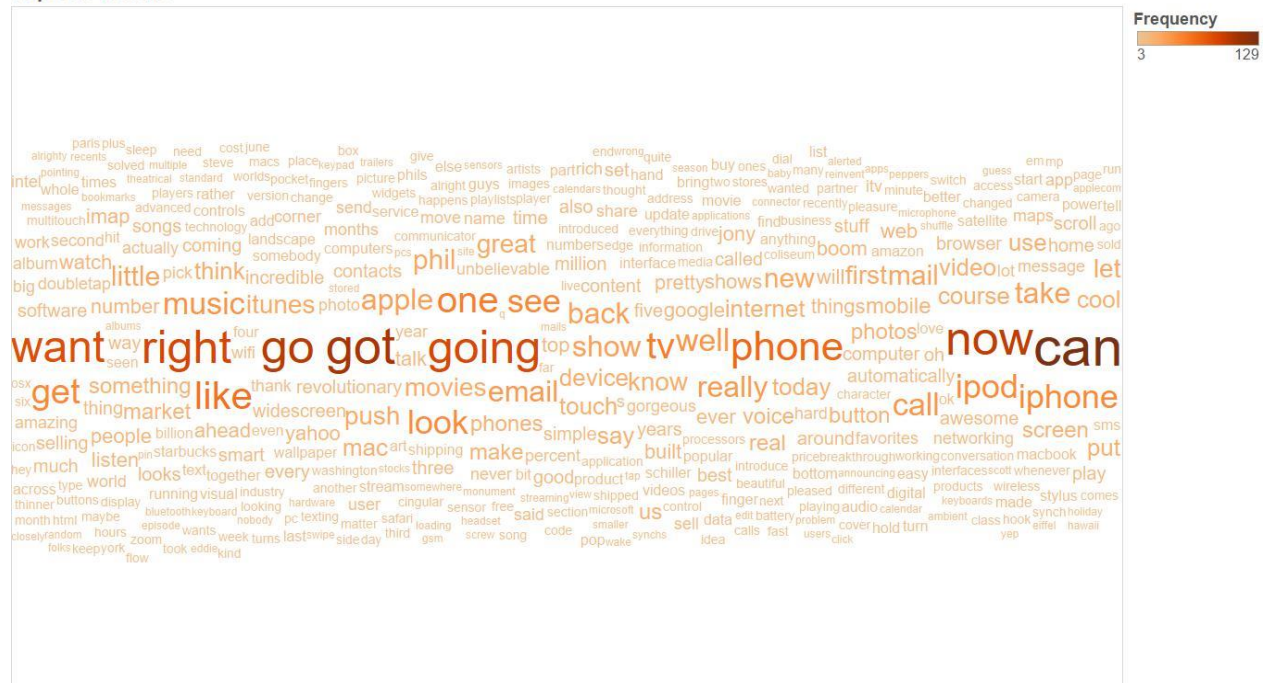
Words. Color shows sum of Frequency. Size shows sum of Frequency.

From those two graphs, I found that the one I set the center as median has more red words than green words. The median is over smaller than the average.



Words. Color shows sum of Frequencyy. Size shows sum of Frequencyy.

Top30%-iPhone



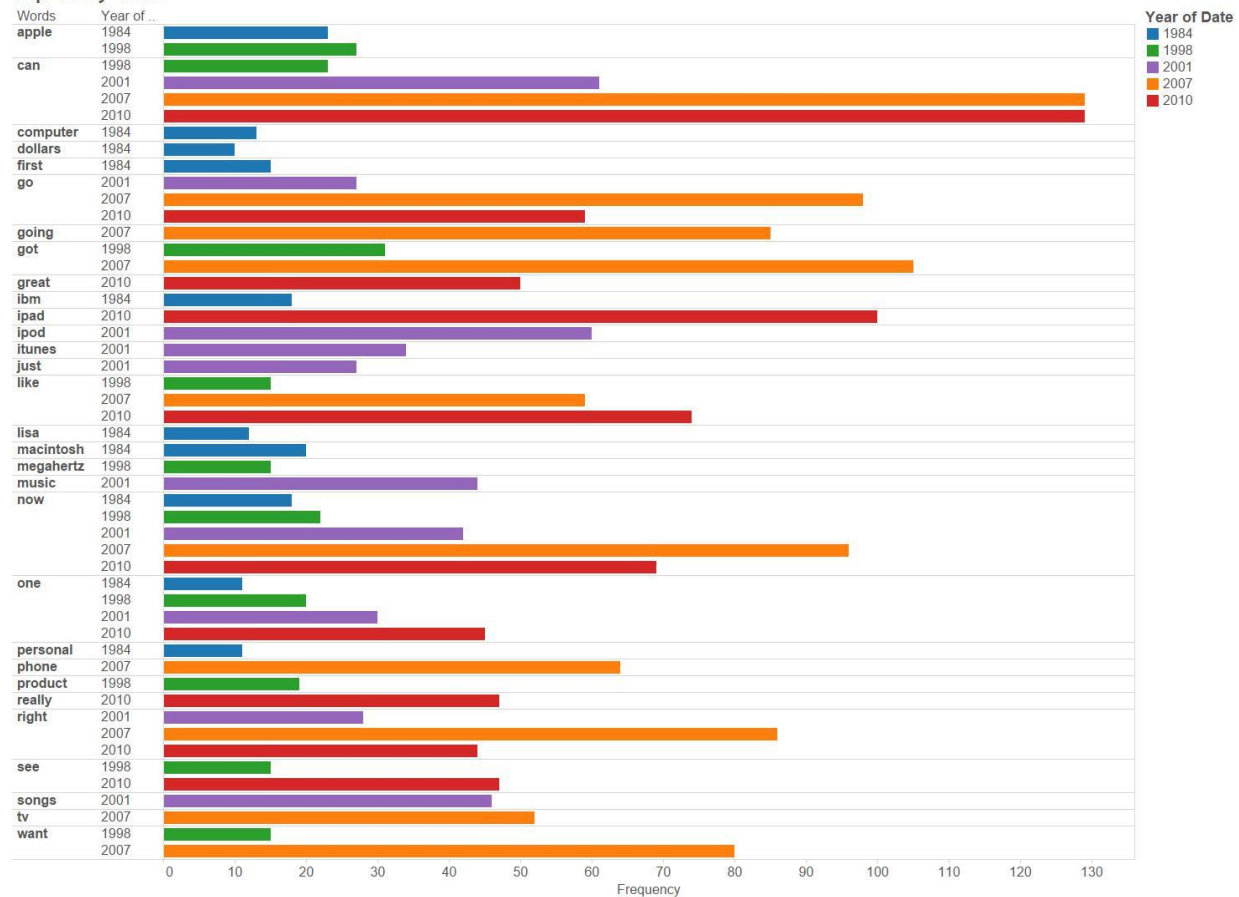
Words. Color shows sum of Frequency. Size shows sum of Frequency.



Words. Color shows sum of Frequency. Size shows sum of Frequency.

By seeing all the images above, I can find big 'now' in all those five word clouds. There are also many other words relevant to time, such as 'today', 'year', in the different places. Let's see the top ten words in each of his speech.

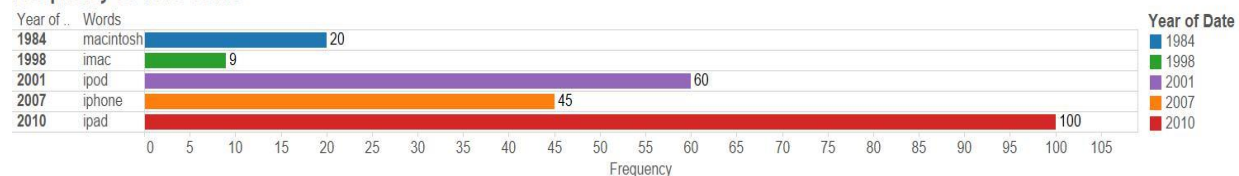
Top-10-By-Years



Sum of Frequency for each Date Year broken down by Words. Color shows details about Date Year.

From the graph above, I find that there are some words only appeared more than once in the top ten words by years. They are 'apple', which is the brand; 'can', 'like', 'want', 'right', show attitudes; 'now', relevant to time; 'go', 'got', 'see', show some directions or achievements. Those words only appeared once are mostly used only for the certain new product in that speech.

Frequency-of-Five-Years



Sum of Frequency for each Words broken down by Date Year. Color shows details about Date Year. The marks are labeled by sum of Frequency. Details are shown for Words. The view is filtered on Words, which keeps imac, ipad, iphone, ipod and macintosh.

Student: Bingyang Dou Subject: CIS576 Assignment 4: Text Analysis/Visualization and Network Visualization

I want to find out how many percent of the products' name he used in his conference, I generated the image for the frequencies and calculated the percentage about the keywords are used: macintosh($20/718=2.78\%$), imac($9/744=1.21\%$), ipod($60/1006=5.96\%$), iphone($45/1388=3.24\%$), ipad($100/1605=6.23\%$). The average of those percentages is 3.884%. The standard deviations is around 3.77%.

Tutorial: <http://georeferenced.wordpress.com/2013/01/15/rwordcloud/> d. References:

<http://onertipaday.blogspot.com/2011/07/word-cloud-in-r.html>

http://www.clearlyandsimply.com/clearly_and_simply/2015/03/word-clouds-withtableau.html